32 Measures of Central Tendency and Dispersion

In this section we discuss two important aspects of data which are its center and its spread. The mean, median, and the mode are **measures of central tendency** that describe where data are centered. The range, variance, and standard deviation are **measures of dispersion** that describe the spread of data.

Measures of Central Tendency: Mode, Median, Mean

Mode
The first measure of the center is the **mode**. It is defined as the value which occurs with the highest frequency in the data. Thus, it is used when one is interested in the most common value in a distribution. A mode may not exist and usually this happens when no data value occurs more frequently than all the others. Also a mode may not be unique.

Example 32.1
The final grades of a class of six graduate students were A, C, B, C, A, B. What is the mode?

Solution.
Since the grades occur at the same frequency then there is no mode for this data.■

Example 32.2
A sample of the records of motor vehicle bureau shows that 18 drivers in a certain age group received $3, 2, 0, 0, 2, 3, 3, 1, 0, 1, 0, 3, 4, 0, 3, 2, 3, 0$ traffic tickets during the last three year. Find the mode?

Solution.
The mode consists of 0 and 3 since they occur six times on the list.■

Median
Another measure of the center is the **median**. The median usually is found when we have an ordered distribution. It is computed as follows. We arrange

the numerical data from smallest to largest. If n denotes the size of the set of data then the median can be found by using the **median rank**

$$MR = \frac{n+1}{2}.$$

If MR is a whole number then the median is the value in that position. If MR ends in .5, we take the sum of the adjacent positions and divide by 2. Unlike the mode, the median always exists and is unique. But it may or may not be one of the given data values. Note that extreme values (smallest or largest) do not affect the median.

Example 32.3
Among groups of 40 students interviewed at each of 10 different colleges, $18, 13, 15, 12, 8, 3, 7, 14, 16, 3$ said that they jog regularly. Find the median.

Solution.
First, arrange the numbers from smallest to largest to obtain

$$3 \quad 3 \quad 7 \quad 8 \quad 12 \quad 13 \quad 14 \quad 15 \quad 16 \quad 18$$

Next, compute the median rank $MR = \frac{10+1}{2} = 5.5$. Hence, the median is $\frac{12+13}{2} = 12.5$■

Example 32.4
Nine corporations reported that in 1982 they made cash donations to $9, 16, 11, 10, 13, 12, 6, 9$, and 12 colleges. Find the median number.

Solution.
Arranging the numbers from smallest to largest to obtain

$$6 \quad 9 \quad 9 \quad 10 \quad 11 \quad 12 \quad 12 \quad 13 \quad 16$$

The median rank is $MR = \frac{9+1}{2} = 5$. The median is 11.■

Arithmetic Mean
Another most widely used measure of the center is the **arithmetic mean** or simply **mean**. The mean of a set of N numbers $x_1, x_2, ..., x_N$, denoted by $\overline{x}$, is defined as

$$\overline{x} = \frac{x_1 + x_2 + ... + x_N}{N}.$$

Unlike the median, the mean can be affected by extreme values since it uses the exact value of each data.

Example 32.5
If nine school juniors averaged 41 on the verbal portion of the PSAT test, at most how many of them can have scored 65 or more?

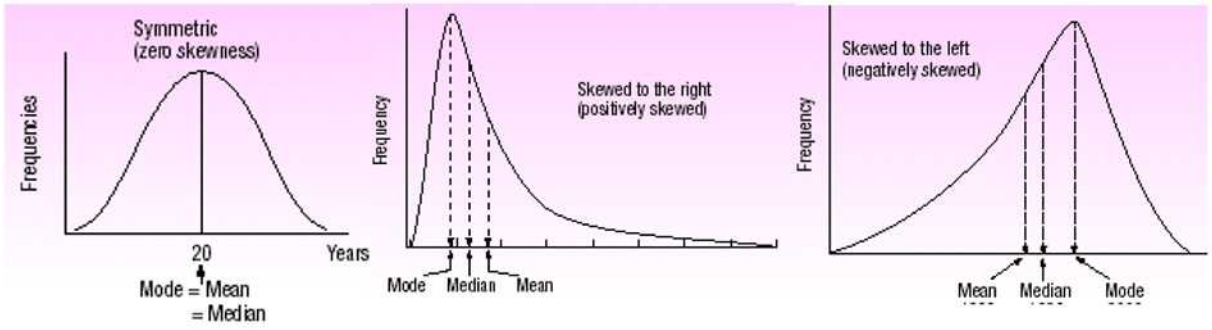
Solution.
We have that $\overline{x} = 41$ and $N = 9$ so that $x_1 + x_2 + \cdots + x_9 = 41 \times 9 = 369$. Since $6 \times 65 = 390 > 369$ and $5 \times 65 = 325$ then at most 5 students can score more than 65.■

Example 32.6
If the numbers $x_1, x_2, ..., x_N$ occur with frequencies $m_1, m_2, ..., m_N$ respectively then what is the mean in this case?

Solution.
The mean is given by

$$\overline{x} = \frac{m_1 x_1 + m_2 x_2 + ... + m_N x_N}{m_1 + m_2 + \cdots + m_N}.$$■

The figure below gives the relationships among the measures of central tendency.



Practice Problems

Problem 32.1
Find (a) the mean, (b) median, and (c) the mode for the following collection of data:

60 60 70 95 95 100

Problem 32.2
Suppose a company employs 20 people. The president of the company earns $200,000, the vice president earns $75,000, and 18 employees earn $10,000 each. Is the mean the best number to choose to represent the "average" salary of the company?

Problem 32.3
Suppose nine students make the following scores on a test:

$$30, 35, 40, 40, 92, 92, 93, 98, 99$$

Is the median the best "average" to represent the set of scores?

Problem 32.4
Is the mode an appropriate "average" for the following test scores?

$$40, 42, 50, 62, 63, 65, 98, 98$$

Problem 32.5
The 20 meetings of a square dance club were attended by 26, 25, 28, 23, 25, 24, 24, 23, 26, 26, 28, 26, 24, 32, 25, 27, 24, 23, 24, and 22 of its members. Find the mode, median, and mean.

Problem 32.6
If the mean annual salary paid to the top of three executives of a firm is $\$96,000$, can one of them receive an annual salary of $\$300,000$?

Problem 32.7
An instructor counts the final examination in a course four times as much as of the four one-hour examinations. What is the average grade of a student who received grades of $74, 80, 61$, and 77 in the four one- hour examinations and 83 in the final examination?

Problem 32.8
In 1980 a college paid its 52 instructors a mean salary of $\$13,200$, its 96 assistant professors a mean salary of $\$15,800$, its 67 associate professors a mean salary of $\$18,900$, and its 35 full professors a mean salary of $\$23,500$. What was the mean salary paid to all the teaching staff of this college?

Measures of Dispersion: Range, Variance, and Standard Deviation
While mean and median tell you about the center of your observations, they say nothing about how data are scattered. **variability** or **dispersion** measures the extent to which data are spread out.
The measures of variability for data that we look at are: the range, the interquartile range, the variance, and the standard deviation.

The Range
To measure the variability between extreme values (i.e. smallest and largest values) one uses the **range**. The range is the difference between the largest and smallest values of a distribution.

Example 32.7
Find the range of each of the following samples:

Sample 1: 6,18,18,18,18,18,18,18,18,18.
Sample 2: 6,6,6,6,6,6,18,18,18,18,18.
Sample 3: 6,7,9,11,12,14,15,16,17,18.

Solution.
Sample 1: 18 - 6 = 12
Sample 2: 18 - 6 = 12
Sample 3: 18 - 6 = 12.■

As you can see from this example, each sample has a range $18 - 6 = 12$ but the spread out of values is quite different in each case. In Sample 1, the spread is uniform whereas it is not in Sample 3. This is a disadvantage of this kind of measure. The range tells us nothing about the dispersion of the values between the extreme (smallest and largest) values. A better understanding is obtained by determining **quartiles**.

Quartiles and Percentiles
Recall that if a set of data is arranged from smallest to largest, the middle value Q_2 (or the arithmetic mean of the two middle values) that divides the set of observations into two equal parts I_1 and I_2 is called the **median** of the distribution. That is, 50 percent of the observations are larger than the median and 50 percent are smaller.

Now, the median of I_1 is denoted by Q_1 (called the **lower quartile**) and that of I_2 by Q_3 (called the **upper quartile**). Thus, Q_1, Q_2 and Q_3 divide the set of data into four equal parts. We call Q_1, Q_2 and Q_3 the three **quartiles** of the distribution.
In a similar manner, one could construct other measures of variation by considering percentiles rather than quartiles. For a whole number P, where $1 \leq P \leq 99$, the Pth **percentile** of a distribution is a value such that $P\%$ of the data fall at or below it. Thus, there are 99 percentiles. The percentiles locations are found using the formula

$$L_P = (n+1) \cdot \frac{P}{100}.$$

Thus, the median Q_2 is the 50th percentile so that $P = 50$ and $L_{50} = \frac{n+1}{2}$ which is the median rank. Note that Q_1 is the 25th percentile and Q_3 is the 75th percentile.
An example will help to explain further.

Example 32.8
Listed below are the commissions earned, in dollars, last month by a sample of 15 brokers at Salomon Smith Barneys office.

2038	1758	1721	1637	2097
2047	2205	1787	2287	1940
2311	2054	2406	1471	1460

(a) Rank the data from smallest to largest.
(b) Find the median rank and then the median.
(c) Find the quartiles Q_1 and Q_3.

Solution.
(a) Arranging the data from smallest to largest we find

1460	1471	1637	1721	1758
1787	1940	2038	2047	2054
2097	2205	2287	2311	2406

(b) The median Q_2 is the 50th percentile so that $P = 50$ and $L_{50} = \frac{15+1}{2} = 8$. Thus, $Q_2 = 2038$.

(c) Q_1 is the 25th percentile so that $L_{25} = (15+1)\cdot\frac{25}{100} = 4$ and so $Q_1 = 1721$. Similarly, Q_3 is the 75th percentile so that $P = 75$ and $L_{75} = (15+1)\cdot\frac{75}{100} = 12$. Thus, $Q_3 = 2205$. ■

Percentiles are extensively used in such fields as educational testing. Undoubtedly some of you have had the experience of being told at what percentile you rated on a scholastic aptitude test.

Example 32.9
You took the English achievement test to obtain college credit in freshman English by examination. If your score was in the 89th percentile, what does this mean?

Solution.
This means 89% of the scores were at or below yours.■

Remark 32.1
In Example 32.8, L_P was found to be a whole number. That is not always the case. For example, if $n = 20$ observations then $L_{25} = (20+1)\cdot\frac{25}{100} = 5.25$. In this case, to find Q_1 we locate the fifth value in the ordered array and then move .25 of the distance between the fifth and sixth values and report that as the first quartile. Like the median, the quartile does not need to be one of the actual values in the data set.
To explain further, suppose a data set contained the six values: 91, 75, 61, 101, 43, and 104. We want to locate the first quartile. We order the values from smallest to largest: 43, 61, 75, 91, 101, and 104. The first quartile is located at

$$L_{25} = (6+1)\cdot\frac{25}{100} = 1.75.$$

The position formula tells us that the first quartile is located between the first and the second value and that it is .75 of the distance between the first and the second values. The first value is 43 and the second is 61. So the distance between these two values is 18. To locate the first quartile, we need to move .75 of the distance between the first and second values, so .75(18) = 13.5. To complete the procedure, we add 13.5 to the first value and report that the first quartile is 56.5.

Box-and-Whisker Plot
A **box-and-whisker plot** is a graphical display, based on quartiles, that

helps us picture a set of data. The steps for making such a box are as follows:

- Draw a vertical (or horizontal) scale to include the lowest and highest data values.
- To the right (or upper for the horizontal scale) of the scale draw a box from Q_1 to Q_3.
- Include a solid line through the box at the median level.
- Draw solid lines, called **whiskers**, from Q_1 to the lowest value and from Q_3 to the largest value.

An example will help to explain.

Example 32.10
Alexanders Pizza offers free delivery of its pizza within 15 km. Alex, the owner, wants some information on the time it takes for delivery. How long does a typical delivery take? Within what range of times will most deliveries be completed? For a sample of 20 deliveries, he determined the following information:

Minimum value	=	13 minutes
Q_1	=	15 minutes
Median	=	18 minutes
Q_3	=	22 minutes
Maximum value	=	30 minutes

Develop a box plot for the delivery times. What conclusions can you make about the delivery times?

Solution.
The box plot is shown in Figure 32.1.

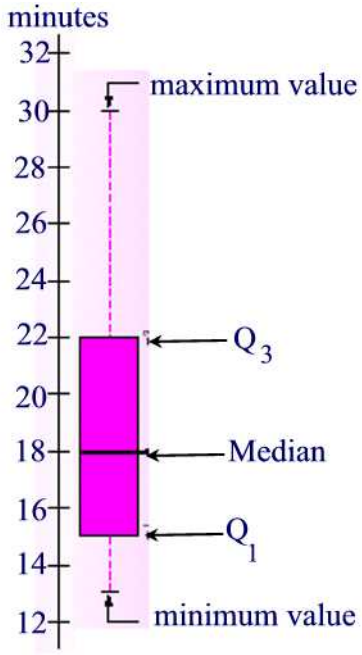


Figure 32.1

The box plot shows that the middle 50 percent of the deliveries take between 15 minutes and 22 minutes. The distance between the ends of the box, 7 minutes, is the **interquartile range**. The interquartile range is the distance between the first and the third quartile. It tells us the spread of the middle half of the data.

The box plot also reveals that the distribution of delivery times is positively skewed since the median is not in the center of the box and the distance from the first quartile to the median is smaller than the distance from the median to the third quartile.

Practice Problems

Problem 32.9

The following table gives the average costs of a single-lens reflex camera:

800	650	300	430	560	470	640	830
400	280	800	410	360	600	310	370

(a) Rank the data from smallest to largest.
(b) Find the quartiles Q_1, Q_2, and Q_3.
(c) Make a box-and-whisker plot.

Problem 32.10
Mr. Eyha took a general aptitude test and scored in the 82nd percentile for aptitude in accounting. What percentage of the scores were at or below his score? What percentage were above?

Problem 32.11
At Center Hospital there is a concern about the high turnover of nurses. A survey was done to determine how long (in months) nurses had been in their current positions. The responses of 20 nurses were:

23	2	5	14	25	36	27	42	12	8
7	23	29	26	28	11	20	31	8	36

(a) Rank the data.
(b) Make a box-and-whisker plot of the data.
(c) What are your conclusions from the plot?

Variance and Standard Deviation
We have seen that a remedy for the deficiency of the range is the use of box and whisker plot. However, an even better measure of variability is the standard deviation. Unlike the range, the variance combines all the values in a data set to produce a measure of spread. The variance and the standard deviation are both measures of the spread of the distribution about the mean. If μ is the population mean of set of data then the quantity $(x - \mu)$ is called the **deviation from the mean**. The **variance** of a data set is the arithmetic average of squared deviation from the mean. It is denoted by s^2 and is given by the formula

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}.$$

Note that the variance is nonnegative, and it is zero only if all observations are the same.
The **standard deviation** is the square root of the variance:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}.$$

The variance is the nicer of the two measures of spread from a mathematical point of view, but as you can see from the algebraic formula, the physical unit of the variance is the square of the physical unit of the data. For example, if our variable represents the weight of a person in pounds, the variance

measures spread about the mean in squared pounds. On the other hand, standard deviation measures spread in the same physical unit as the original data, but because of the square root, is not as nice mathematically. Both measures of spread are useful.

A step by step approach to finding the standard deviation is:

1. Calculate the mean.
2. Subtract the mean from each observation.
3. Square each result.
4. Add these squares.
5. Divide this sum by the number of observations.
6. Take the positive square root.

The variance and standard deviation introduced above are for a population. We define the **sample variance** by the formula

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

and the **sample standard deviation** by the formula

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}.$$

The reason that for s we use $n-1$ instead of n is because usually a sample does not contain extreme values and we want s to be an estimate of σ therefore by using $n-1$ we make s a little larger than if we divide by n.

Example 32.11
The owner of the Ches Tahoe restaurant is interested in how much people spend at the restaurant. He examines 10 randomly selected receipts for parties of four and writes down the following data.

44 50 38 96 42 47 40 39 46 50

(a) Find the arithmetic mean.
(b) Find the variance and the standard deviation.

Solution.
(a) The arithmetic mean is the sum of the above values divided by 10,i.e., $\overline{x} = 49.2$
Below is the table for getting the standard deviation:

x	$x - 49.2$	$(x - 49.2)^2$
44	-5.2	27.04
50	0.8	0.64
38	11.2	125.44
96	46.8	2190.24
42	-7.2	51.84
47	-2.2	4.84
40	-9.2	84.64
39	-10.2	104.04
46	-3.2	10.24
50	0.8	0.64
Total		2600.4

Hence the variance is $s^2 = \frac{2600.4}{9} \approx 289$ and the standard deviation is $s = \sqrt{289} = 17$. What this means is that most of the patrons probably spend between 49.20 - 17 =\$32.20 and 49.20+17=\$66.20. ■

Practice Problems

Problem 32.12
The following are the wind velocities reported at 6 P.M.on six consecutive days: 13,8,15,11,3 and 10. Find the range, sample mean, sample variance, and sample standard deviation.

Problem 32.13
An airline's records show that the flights between two cities arrive on the average 4.6 minutes late with a standard deviation of 1.4 minutes. At least what percentage of its flights between these two cities arrive anywhere between 1.8 minutes late and 7.4 minutes late?

Problem 32.14
One patient's blood pressure, measured daily over several weeks, averaged 182 with a standard deviation of 5.3, while that of another patient averaged 124 with a standard deviation of 9.4. Which patient's blood pressure is relatively more variable?

Problem 32.15
By sampling different landscapes in a national park over a 2-year period, the number of deer per square kilometer was determined. The results were (deer per square kilometer)

30	20	5	29	58	7
20	18	4	29	22	9

Compute the range, sample mean, sample variance, and sample standard deviation.

Problem 32.16
A researcher wants to find the number of pets per household. The researcher conducts a survey of 35 households. Find the sample variance and standard deviation.

0	2	3	1	0
1	2	3	1	0
1	2	1	1	0
3	2	1	1	1
4	1	2	2	4
3	2	1	2	3
2	3	4	0	2

Problem 32.17
Suppose two machines produce nails which are on average 10 inches long. A sample of 11 nails is selected from each machine.

Machine A: 6, 8, 8, 10, 10, 10, 10, 10, 12, 12, 14.
Machine B: 6, 6, 6, 8, 8, 10, 12, 12, 14, 14, 14.

Which machine is better than the other?

Problem 32.18
Find the missing age in the following set of four student ages.

Student	Age	Deviation from the Mean
A	19	-4
B	20	-3
C	?	1
D	29	6

Problem 32.19
The maximum heart rates achieved while performing a particular aerobic exercise routine are measured (in beats per minute) for 9 randomly selected individuals.

145 155 130 185 170 165 150 160 125

(a) Calculate the range of the time until failure.
(b) Calculate the sample variance of the time until failure.
(c) Calculate the sample standard variation of the time until failure.

Problem 32.20
The following data gives the number of home runs that Babe Ruth hit in each of his 15 years with the New York Yankees baseball team from 1920 to 1934:

54 59 35 41 46 25 47 60 54 46 49 46 41 34 22.

The following are the number of home runs that Roger Maris hit in each of the ten years he played in the major leagues from 1957 on:

8 13 14 16 23 26 28 33 39 61

Calculate the mean and standard deviation for each player's data and comment on the consistency of performance of each player.

Problem 32.21
An office of Price Waterhouse Coopers LLP hired five accounting trainees this year. Their monthly starting salaries were: \$2536; \$2173; \$2448; \$2121; and \$2622.

(a) Compute the population mean.
(b) Compute the population variance.
(c) Compute the population standard deviation.

Normal Distribution
To better understand how standard deviations are used as measures of dispersion, we next consider normal distributions. The graph of a normal distribution is called a **normal curve** or a **bell-shaped curve.** The curve has the following properties:

1. The curve is bell-shaped with the highest point over the mean μ.
2. It is symmetrical about the line through μ.
3. The curve approaches the horizontal axis but never touches or crosses it.
4. The points where the curve changes concavity occur at $\mu+\sigma$ and $\mu-\sigma$.
5. The total area under the curve is assumed to be 1.(0.5 to the left of the mean and 0.5 to the right).

The data for a normal distribution are spread according to the following rules (See Figure 32.2):

- About 68 percent of the data values will lie within one standard deviation of the mean.
- About 95 percent of the data values will lie within two standard deviations of the mean.
- About 99.7 percent of the data values will lie within three standard deviations of the mean.

This result is sometimes referred to as the **emperical rule**, because the given percentages are observed in practice.

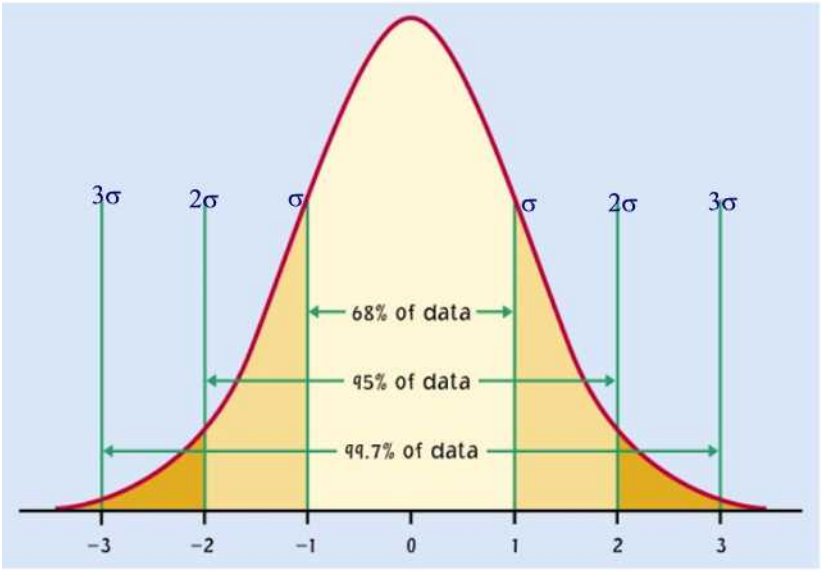


Figure 32.2

Example 32.12
When a standardized test was scored, there was a mean of 500 and a standard deviation of 100. Suppose that 10,000 students took the test and their scores had a bell-shaped distribution.

(a) How many scored between 400 and 600?
(b) How many scored between 300 and 700?
(c) How many scored between 200 and 800?

Solution.
(a) Since one standard deviation on either side of the mean is from 400 to 600, about 68% of the scores fall in this interval. Thus, $0.68 \times 10,000 = 6800$ students scored between 400 and 600.
(b) About 95% of 10,000 or 9500 students scored between 300 and 700.
(c) About 99.7% of 10,000 or 9970 students scored between 200 and 800■

By the discussion above we get information of the percentage of data within a certain number of standard deviations. However, if we want to find the location of a data value x from the mean we can use the so-called **z-score** given by the formula

$$z = \frac{x - \mu}{\sigma}$$

Now, if the z-score is given then the raw data can be found by solving the equation $z = \frac{x-\mu}{\sigma}$ for x to obtain

$$x = \mu + z\sigma.$$

Note that from this formula we see that the value of z tells us how many standard deviation the corresponding value of x lies above (if $z > 0$) or below (if $z < 0$) the mean of its distribution.

Example 32.13
Scores on intelligence tests (IQs) are normally distributed in children. IQs from the Wechsler intelligence tests are known to have means of 100 and standard deviations of 15. In almost all the states in the United States (Pennsylvania and Nebraska are exceptions) children can be labeled as mentally retarded if their IQ falls to 70 points or below. What is the maximum z score one could obtain on an intelligence test and still be considered to be mentally retarded?

Solution.
Applying the z-score formula we find $z = \frac{70-100}{15} = -2$.■

Example 32.14
In a certain city the mean price of a quart of milk is 63 cents and the standard deviation is 8 cents. The average price of a package of bacon is \$1.80 and the standard deviation is 15 cents. If we pay \$0.89 for a quart of milk and \$2.19 for a package of bacon at a 24-hour convenience store, which is relatively more expensive?

Solution.
To answer this, we compute z-scores for each:

$$z_{Milk} = \frac{0.89 - 0.63}{0.08} = 3.25$$

and

$$z_{Bacon} = \frac{2.19 - 1.80}{0.15} = 2.60.$$

Our z-scores show us that we are overpaying quite a bit more for the milk than we are for the bacon.■

Example 32.15
Graduate Record Examination (GRE) scores have means equal to 500 and standard deviations of 100. If a person receives a z-score on the GRE of 1.45, what would their raw score be?

Solution.
Using the formula $x = \mu + z\sigma$ we find $x = 500 + 1.45(100) = 645$.■

Practice Problems

Problem 32.22
On a final examination in Statistics, the mean was 72 and the standard deviation was 15. Assuming normal distribution, determine the z-score of students receiving the grades (a) 60, (b) 93, and (c) 72.

Problem 32.23
Referring to the previous exercise, find the grades corresponding to the z-score $z = 1.6$.

Problem 32.24
If $z_1 = 0.8$, $z_2 = -0.4$ and the corresponding x-values are $x_1 = 88$ and $x_2 = 64$ then find the mean and the standard deviation, assuming we have a normal distribution.

Problem 32.25
A student has computed that it takes an average (mean) of 17 minutes with a standard deviation of 3 minutes to drive from home, park the car, and walk to an early morning class. Assuming normal distribution,

(a) One day it took the student 21 minutes to get to class. How many standard deviations from the average is that?
(b) Another day it took only 12 minutes for the student to get to class. What is this measurement in standard units?
(c) Another day it took him 17 minutes to be in class. What is the z-score?

Problem 32.26
Mr. Eyha's z-score on a college exam is 1.3. If the x-scores have a mean of 480 and a standard deviation of 70 points, what is his x-score?

Problem 32.27
(a) If $\mu = 80, \sigma = 10$, what is the z-score for a person with a score of 92?
(b) If $\mu = 65, \sigma = 12$, what is the raw score for a z-score of -1.5?

Problem 32.28
Sketch a normal curve. Mark the axis corresponding to the parameter μ and the axis corresponding to $\mu + \sigma$ and $\mu - \sigma$.

Problem 32.29
For the population of Canadian high school students, suppose that the number of hours of TV watched per week is normally distributed with a mean of 20 hours and a standard deviation of 4 hours. Approximately, what percentage of high school students watch

(a) between 16 and 24 hours per week?
(b) between 12 and 28 hours per week?
(c) between 8 and 32 hours per week?

Problem 32.30
The length of human pregnancies from conception to birth varies according to a distribution that is approximately normal with mean 266 days and standard deviation 16 days. Use the emperical rule to answer the following questions.

(a) Between what values do the lengths of the middle 95% of all pregnancies fall?
(b) How short are the shortest 2.5% of all pregnancies?